CLAIMS

Sub a'

1. In a cluster of computing nodes having shared access to one or more volumes of data storage using a parallel file system, a method for managing the data storage, comprising:

initiating a session of a data management (DM) application on a first one of the nodes;

running a user application on a second one of the nodes;

receiving a request submitted to the parallel file system by the user application on the second node to perform a file operation on a file in one of the volumes of data storage; and

sending a DM event message from the second node to the first node responsive to the request, for processing by the data management application on the first node.

2. A method according to claim 1, wherein initiating the session comprises initiating the session in accordance with a data management application programming interface (DMAPI) of the parallel file system, and wherein receiving the request comprises processing the request using the DMAPI.

3. A method according to claim 2, and comprising receiving and processing the event message at the first node using one or more functions of the DMAPI called by the data management application.

4. A method according to claim 2, wherein sending the event message comprises sending the message for processing in accordance with a disposition specified by the data management application using the DMAPI for

association with an event generated by the file operation.

5. A method according to claim 1, and comprising:

receiving a response to the event message from the data management application on the first node; and

performing the file operation requested by the user application on the second node subject to the response from the data management application on the first node.

6. A method according to claim 5, wherein receiving the request comprises submitting the request using a file operation thread running on the second node, and blocking the thread until the response to the event message is received from the first node.

7. A method according to claim 5, wherein sending the event message comprises passing the event message from a source physical file system (PFS) on the second node to a session PFS on the first node, and wherein receiving the response comprises passing a response message from the session PFS to the source PFS.

8. A method according to claim 1, and comprising:

receiving the event message at the first node;

obtaining a data management access right from a physical file system (PFS) at the first node responsive to the event message; and

processing the event message using the access right.

9. A method according to claim 1, wherein receiving the request comprises receiving first and second requests of different types submitted to a physical file system (PFS) at the second node, and wherein based on the different request types, sending the event message comprises

sending a first event message to the first node responsive to the first request, and sending a second event message responsive to the second request to a further node, on which a further data management application session has been initiated.

10. A method according to claim 9, wherein sending the first and second event messages comprises:

receiving at the second node a specification of event types and their respective dispositions, the event types corresponding to the requests to perform the file operations, and dispositions indicating which of the event messages should be sent to which of the nodes; and

sending the messages responsive to the specification.

11. A method according to claim 1, wherein running the user application comprises running a first user application instance on the second node, and running a further user application instance on a further one of the nodes, and comprising receiving a further request submitted to the parallel file system by the further user application instance to perform a further file operation, and sending a further event message responsive to the further request for processing by the data management application on the first node.

12. A method according to claim 11, wherein the further one of the nodes is the first node.

13. A method according to claim 1, wherein initiating the session of the data management application comprises initiating a data migration application, so as to free storage space on at least one of the volumes of data storage.

14. A method according to claim 1, and comprising choosing one of the nodes to act as a session manager node, wherein initiating the session comprises sending a message from the session node to the session manager node, causing the session manager node to distribute a specification of events and respective dispositions of the events for the session among the nodes in the cluster, and wherein sending the DM event message comprises sending the message in accordance with the dispositions.

15. A method according to claim 14, wherein one of the nodes is appointed to serve as a respective file system manager for each of one or more file systems in the cluster, and wherein for each of the file systems, the session manager node sends the specification of the dispositions applicable to the file system to the respective file system manager, which sends the dispositions to all of the nodes in the cluster on which the file system is mounted.

16. A method according to claim 1, wherein sending the DM event message comprises incorporating in the message a data field uniquely identifying the second node.

17. A method according to claim 1, and comprising receiving from one of the nodes other than the first one of the nodes a call for a data management application programming interface (DMAPI) function in connection with the session, and performing the function only if it does not change a state of the session or of an event associated with the session.

18. Computing apparatus, comprising:

one or more volumes of data storage, arranged to store data; and

a plurality of computing nodes, linked to access the volumes of data storage using a parallel file system, and arranged so as to enable a data management (DM) application to initiate a data management session on a first one of the nodes, while allowing a user application to run on a second one of the nodes, so that when the user application submits a request to the parallel file system on the second node to perform a file operation on a file in one of the volumes of data storage, a DM event message is sent from the second node to the first node responsive to the request, for processing by the data management application on the first node.

19. Apparatus according to claim 18, wherein the session is initiated in accordance with a data management application programming interface (DMAPI) of the parallel file system, and wherein the request is processed using the DMAPI.

20. Apparatus according to claim 19, and wherein the event message is received and processed at the first node using one or more functions of the DMAPI called by the data management application.

21. Apparatus according to claim 19, wherein the event message is sent for processing in accordance with a disposition specified by the data management application using the DMAPI for association with an event generated by the file operation.

22. Apparatus according to claim 17, wherein the nodes are arranged so that the data management application on

the first node generates a response to the event message, and the file operation requested by the user application is performed on the second node subject to the response from the data management application on the first node.

23. Apparatus according to claim 22, wherein the request is submitted using a file operation thread running on the second node, and the thread is blocked until the response to the event message is received from the first node.

24. Apparatus according to claim 22, wherein the event message is passed from a source physical file system (PFS) on the second node to a session PFS on the first node, and wherein the response comprises a response message passed from the session PFS to the source PFS.

25. Apparatus according to claim 18, wherein when the event message is received at the first node, a data management access right is obtained from the physical file system (PFS) at the first node responsive to the event message, and the event message is processed using the access permission.

26. Apparatus according to claim 17, wherein when first and second file operation requests of different types are submitted to the physical file system (PFS) at the second node, and wherein based on the different request types, the second node is arranged to send a first event message to the first node responsive to the first request, and a second event message responsive to the second request to a further node, on which a further data management application session has been initiated.

27. Apparatus according to claim 26, wherein the first and second event messages are sent after receiving at the

second node a specification of event types and their respective dispositions, the event types corresponding to the requests to perform the file operations, and dispositions indicating which of the event messages should be sent to which of the nodes, such that the second node sends the messages responsive to the specification.

28. Apparatus according to claim 18, wherein the user application comprises a first user application instance running on the second node, and a further user application instance running on a further one of the nodes, wherein responsive to a further request submitted to the parallel file system by the further user application instance to perform a further file operation, a further event message responsive to the further request is sent for processing by the data management application on the first node.

29. Apparatus according to claim 28, wherein the further one of the nodes is the first node.

30. Apparatus according to claim 18, wherein the data management application comprises a data migration application, for freeing storage space on at least one of the volumes of data storage.

31. Apparatus according to claim 18, wherein one of the nodes is chosen to act as a session manager node, wherein the session is initiated by sending a message from the first node to the session manager node, causing the session manager node to distribute a specification of events and respective dispositions of the events for the session among the nodes in the cluster, and wherein the

DM event message is sent in accordance with the dispositions.

32. Apparatus according to claim 18, wherein one of the nodes is appointed to serve as a respective file system manager for each of one or more file systems in the cluster, and wherein for each of the file systems, the session manager node is arranged to send the specification of the dispositions applicable to the file system to the respective file system manager, which sends the dispositions to all of the nodes in the cluster on which the file system is mounted.

33. Apparatus according to claim 18, wherein the second node is arranged to incorporate in the DM message a data field uniquely identifying the second node.

34. Apparatus according to claim 18, wherein upon receiving from one of the nodes other than the first one of the nodes a call for a data management application programming interface (DMAPI) function in connection with the session, the nodes are arranged to perform the function only if it does not change a state of the session or of an event associated with the session.

35. A computer software product for use in a cluster of computing nodes having shared access to one or more volumes of data storage using a parallel file system, the product comprising a computer-readable medium in which program instructions are stored, which instructions, when read by the computing nodes, cause a session of a data management (DM) application to be initiated on a first one of the nodes, while allowing a user application to run on a second one of the nodes, and in response to a

request submitted to the parallel file system by the user application on the second node to perform a file operation on a file in one of the volumes of data storage, cause the second node to send a DM event message to the first node, for processing by the data management application on the first node.

36. A product according to claim 35, wherein the product comprises a data management application programming interface (DMAPI) of the parallel file system, and wherein the request is processed using the DMAPI.

37. A product according to claim 36, and wherein the event message is received and processed at the first node using one or more functions of the DMAPI called by the data management application.

38. A product according to claim 36, wherein the event message is sent for processing in accordance with a disposition specified by the data management application using the DMAPI for association with an event generated by the file operation.

39. A product according to claim 35, wherein the instructions cause the data management application on the first node to generate a response to the event message, whereupon the file operation requested by the user application is performed on the second node subject to the response from the data management application on the first node.

40. A product according to claim 39, wherein the request is submitted using a file operation thread running on the second node, and the thread is blocked until the response to the event message is received from the first node.

41. A product according to claim 39, wherein the event message is passed from a source physical file system (PFS) on the second node to a session PFS on the first node, and wherein the response comprises a response message passed from the session PFS to the source PFS.

42. A product according to claim 35, wherein when the event message is received at the first node, a data management access right is obtained from the physical file system (PFS) at the first node responsive to the event message, and the event message is processed using the access permission.

43. A product according to claim 35, wherein first and second file operation requests of different types are submitted to the physical file system (PFS) at the second node, and wherein based on the different request types, the instructions cause the second node to send a first event message to the first node responsive to the first request, and a second event message responsive to the second request to a further node, on which a further data management application session has been initiated.

44. A product according to claim 43, wherein the first and second event messages are sent after receiving at the second node a specification of event types and their respective dispositions, the event types corresponding to the requests to perform the file operations, and dispositions indicating which of the event messages should be sent to which of the nodes, such that the second node sends the messages responsive to the specification.

45. A product according to claim 35, wherein the user application comprises a first user application instance running on the second node, and a further user application instance running on a further one of the nodes, wherein responsive to a further request submitted to the parallel file system by the further user application instance to perform a further file operation, a further event message responsive to the further request is sent for processing by the data management application on the first node.

46. A product according to claim 45, wherein the further one of the nodes is the first node.

47. A product according to claim 35, wherein the data management application comprises a data migration application, for freeing storage space on at least one of the volumes of data storage.

48. A product according to claim 35, wherein the instructions cause one of the nodes to be chosen to act as a session manager node, and wherein the session is initiated by sending a message from the first node to the session manager node, causing the session manager node to distribute a specification of events and respective dispositions of the events for the session among the nodes in the cluster, and wherein the DM event message is sent in accordance with the dispositions.

49. A product according to claim 35, wherein one of the nodes is appointed to serve as a respective file system manager for each of one or more file systems in the cluster, and wherein for each of the file systems, the instructions cause the session manager node to send the

specification of the dispositions applicable to the file system to the respective file system manager, which sends the dispositions to all of the nodes in the cluster on which the file system is mounted.

50. A product according to claim 35, wherein the instructions cause the second node to incorporate in the DM message a data field uniquely identifying the second node.

51. A product according to claim 35, wherein upon receiving from one of the nodes other than the first one of the nodes a call for a data management application programming interface (DMAPI) function in connection with the session, the instructions cause the nodes to perform the function only if it does not change a state of the session or of an event associated with the session.